Pattern-based English-Latvian Toponym Translation

The paper on similar issues will be presented at European Association for Machine Translation conference.

Tatiana Gornostay
Tilde, Latvia
tatjana.gornostaja@tilde.lv

Inguna Skadiņa
Tilde, Latvia
inguna.skadina@tilde.lv

Abstract

Due to their linguistic and extra-linguistic nature toponyms deserve a special treatment when they are translated. The paper deals with issues related to automated translation of toponyms from English into Latvian. Translation process allows us to translate not only toponyms from a dictionary, but out-of-vocabulary toponyms as well. Translation of out-of-vocabulary toponyms is divided into three steps: source string normalization, translation, and target string normalization. Translation step implies application of translation strategies and linguistic toponym translation patterns. 10,000 UK-related toponyms from Geonames were used as a development set. The developed methods have been evaluated on a test set: the accuracy of translation is 67% for the whole test set, 58% for one-word toponymic units, and 81% for multi-word toponyms.

1 Introduction

Toponyms in general are studied by toponymy, they represent names of places comprising the following types:

- hydronyms (names of bodies of water: bays, streams, lakes, lagoons, oceans, ponds, seas, etc.);
- oronyms (names of mountains, cliffs, craters, rocks, points, etc.);
- geonyms (general names for streets, squares, lines, avenues, paths, alleys, roads, embankments, etc.);
- oeconyms (names of populated places: an administrative division, country, city, town, house or other building);
- cosmonyms or astronyms (names of stars, constellations or other heavenly bodies).

The paper aims to research a complicated task of machine translation (MT) and cross-language information retrieval (CLIR) – automated translation of toponyms. Most of toponym translation approaches are data-driven (see, e.g. Meng et al., 2001; Al-Onaizan and Knight, 2002; Sproat et al., 2006; Alegria et al., 2006; Wentland et al., 2008) since they deal with widely used languages which have enough linguistic resources for development.

Taking into account an under-resourced status of the Latvian language with few available corpus resources, especially parallel bilingual corpora, a rule-based approach is proposed for the English-Latvian toponym translation.

There are several commonly used translation strategies for toponyms (Babych and Hartley, 2004): transference strategy (i.e., do-not-translate), transliteration strategy (i.e., phonetic or spelling rendering), translation strategy (i.e., translation itself) and combined strategy.

Transference strategy with a do-not-translate list is often used for translation of toponyms which do not need any rendering at all and are often left not translated, e.g. organization names (Babych and Hartley, 2003) or names of hotels in our system.

The most common transliteration techniques are phoneme-based and grapheme-based (Zhang et al., 2004). The phoneme-based approach (Knight and

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 41–47

Graehl, 1998; Meng et al., 2001; Oh and Choi, 2002; Lee and Chang, 2003) implies conversion of a source language word into a target language word via its phonemic representation, i.e., grapheme-phoneme-grapheme conversion. The grapheme-based technique converts a source language word into a target language word without any phonemic representation (grapheme-grapheme conversion) (Stalls and Knight, 1998; Li et al., 2004).

The first part of the paper presents an overview of the concept and nature of toponyms. In the second part we focus on the English-Latvian toponym translation, including the description of translation strategies (TS) and linguistic toponym translation patterns (LTTP).

2 Concept and Nature of Toponyms

Although Geoffrey Leech (1981) accepts a special status of toponyms as proper names without a conceptual meaning since any componential analysis cannot be performed for them, we should bear in mind and admit the fact that many toponyms are at least meaningful etymologically, e.g *Cambridge* – bridge over the river *Cam* (Leidner, 2007).

Toponyms are also ambiguous. Leidner (2007) describes three types of toponymical ambiguity:

- morpho-syntactic ambiguity: a word itself may be a toponym or may be a non-toponym, e.g. *Liepa* as a populated place in Latvia versus *liepa* (lime-tree) as a common noun;
- referential ambiguity: a toponym may refer to more than one place of the same type, e.g. *Riga* as a populated place and the capital of Latvia and *Riga* as a populated place in the USA, state Michigan;
- feature type ambiguity: a toponym may refer to more than one place of a different type, e.g. *Ogre* as a populated place and a river in Latvia.

Another type of toponymical ambiguity is eponymical ambiguity when places are named after people or deities, e.g., *Vancouver* after George Vancouver. Sometimes the same place is known by different names – endonyms (names of places used by inhabitants, self-assigned names) and exonyms (names of places used by other groups, not locals), e.g. *Firenze* for its inhabitants and *Florence* for English.

Furthermore, metonymy also contributes to the issue. This linguistic phenomenon was studied from the toponymical point of view by Markert and Nissim (2002). The authors stated that metonymic use of toponyms is regular and productive. It can reach up to 17% of all of toponyms as it was proved by the example of the English language. The most frequent and conventional case of toponymical metonymy is as in the "*government of …*" pattern, e.g. "*Latvia announced …*" means "*the government of Latvia announced …*".

Finally, toponyms are changed frequently since they themselves and the places they refer to are not constant. Therefore, when dealing with toponyms it is also very important to take into consideration historical and cultural facts.

Thus, the abovementioned linguistic and extra-linguistic features make toponym processing difficult, i.e., their resolution, retrieval, and especially translation.

3 English-Latvian Toponym Translation

In the overall MT, English-Latvian toponym translation problems have not been researched in before. The existing literature describes general principles of rendering of the English proper names, mostly anthroponyms, into Latvian. Therefore we studied three main issues related to MT of the English-Latvian toponyms:

- orthographic, phonetic and grammatical distinctions between these languages;
- potential toponym translation strategies;
- potential linguistic toponym translation patterns.

Although English and Latvian are Indo-European languages and share some grammatical features, they have a lot of differences. At first, English belongs to the Germanic language group while Latvian belongs to the group of the Baltic languages. In morphological typology the English language is an analytical language in contrast to a synthetic Latvian with a rich set of inflections.

The linguistic features of Latvian toponymic units were studied to ensure that translations correspond to common rules of the Latvian grammar and orthography. For instance, Latvian multi-word

units can be translated in several ways, however, a compound is preferable if the source toponymic unit could be reconstructed (Ahero, 2006).

The lack of orthographic and phonetic convergence in English (26 letters to 44 phonemes), historical changes and traditions in spelling, origin language of a toponym, and ambiguity were the main difficulties we faced.

3.1 Source String Normalization

The process of translation of a toponymic unit is divided into three steps: source string normalization, translation, i.e., application of translation strategy (TS) and linguistic toponym translation patterns (LTTP), and target string normalization according to the Latvian grammar and orthography rules.

Source string normalization implies the following changes:

- all tabs and double space characters, including the string beginning, are normalized to single space characters;
- the so-called "zero-fertility words" (Al-Onaizan and Knight, 2002) of English are normalized to zero-translations into Latvian, e.g. the indefinite article *a* is omitted;
- hyphenated words are replaced with non-hyphenated ones;
- some abbreviations are expanded to full words, e.g. *St.* to *Saint*;
- signs, if possible, are replaced with words, e.g. *&* to *and*;
- punctuation marks are normalized to zero-translations.

3.2 Translation: English-Latvian Toponym Translation Strategies

The English-Latvian transliteration strategy is based on the grapheme-to-grapheme approach, which implies direct mapping of English letter sequences into Latvian ones, formalized in a set of transliteration rules. Transliteration strategy is language dependent (Karimi et al., 2007). It is not a trivial task, due to issues described above, as well as due to many exceptions (see Castañeda-Hernández, 2004 about general toponym translation problem).

The set of English-Latvian transliteration rules consists of about 110 transliteration patterns describing English-Latvian grapheme-to-grapheme correspondences. All foreign names (those of non-English origin) are rendered according to English pronunciation standards. The main principle is the possibility to reconstruct the source toponymic unit (Ahero, 2006).

The result of transliteration may vary, as there are several ways of rendering English letter combinations into Latvian, e.g., *-c-* stands for *-k-* before consonants (except *-h-*), and *-a-*, *-o-*, *-u-*, for *-s-* before *-i-*, *-e-*, *-y-*, and for *-č-* in the combination with *-h-*.

Transference strategy is applied to both unprocessed toponymic units, which are not described by any of linguistic toponym translation patterns, and organization and hotel names.

There are cases when multi-word toponyms are not transferred or transliterated but translated into Latvian, e.g., *East Anglian Heights*, *North West Highlands* are translated into Latvian as *Austrumanglijas augstiene, Ziemeļskotijas kalnāji* correspondingly. Single word units are transliterated, as a rule.

Transliteration strategy can be also applied to multi-word units in parallel with translation which is infrequent and conventional.

Toponym translation strategies are closely related with LTTPs and are language dependent. Therefore combined strategy is also used when treating different types of toponyms.

3.3 Translation: Linguistic Toponym Translation Patterns

Most of popular toponyms, such as names of countries and capitals, seas and oceans, are translated using an English-Latvian dictionary, e.g., *Lisbon – Lisabona*, *Brussels – Brisele*, *Cologne – Ķelne*, *Antwerp – Antverpene*, *Great Britain – Lielbritānija*, *Atlantic Ocean – Atlantijas okeāns*. If a toponym is an out-of-vocabulary (OOV) word then one of the LTTPs is applied.

To determine common LTTPs for toponyms which are not in dictionaties we used a list of 10,000 UK-related toponyms from Geonames and analyzed 59 most common toponym types.

LTTPs determine ways how source toponymic units are rendered into target toponymic units. We distinguish two types of LTTPs: in-word patterns and multi-word patterns.

The in-word LTTP describes word transformation model based on English-Latvian transliteration rules, including the most frequent prefixes, suffixes, and letter combinations. There are about 300 in-word LTTPs described, e.g.: *new-* to *ņū-*, *deep-* to *dīp-*, *mc-* to *mak-*, *-worth* to *–vērt, -islet* to *–ailet*, etc.

Multi-word LTTPs involve three translation strategies. The first translation strategy S_1 is based on transliteration rules. Translation strategy S_2 combines the translation strategy S_1 with the insertion of a nomenclature word, e.g., *Bebington* (as a railroad station) – *Bebingtonas stacija*. If a nomenclature word is included in a source toponymic unit, as it is in the pattern S_3, it is either translated (*Newton Point - Ņūtona zemesrags, Gog Magog Hills - Gogmagogu kalni*) or transliterated (*Green Isle – Grīnaila, North East Coast – Nortīstkosta*) in the target language.

We have described 40 nomenclature words which are translated under certain conditions. Auxiliary words, such as prepositions, are also either translated or transliterated, e.g., *Horse of Copinsay – Horsofkopinsejs* (transliteration), *Milford upon Sea - Milforda pie jūras* (translation).

Examples of LTTPs are presented in Table 1. X_n is a toponymic unit in a source language, S_n is a translation strategy, Y_n is a toponymic unit in a target language, and $P_n\{X_n, S_n, Y_n\}$ is a corresponding LTTP.

3.4 Target String Normalization

Target string normalization modifies a toponymic unit according to the Latvian grammar and orthography rules, e.g. all populated places are feminine gender (see P2): *Newcastle → Ņūkāsla* which is indicated by the ending *–a* (feminine, singular nominative).

English Toponym X_n	**Translation Pattern P_n**	**Translation Strategy S_n**	**Latvian Toponym Y_n**
$P_1\{X_1, S_1, Y_1\}$			
X1: N *Knocklayd*	P1: N → N	S1: transliteration	Y1: N masculine singular *Nokleids*
$P_2=\{X_1, S_1, Y_2\}$			
X1: N *Newcastle*	P2: N → N	S1: transliteration	Y2: N feminine singular *Ņūkāsla*
$P_3=\{X_1, S_2, Y_3\}$			
X1: N *Bebington*	P3: N → N + N	S2: transliteration + nomenclature word	Y3: N feminine singular genitive + N *Bebingtonas stacija*
$P_4=\{X_2, S_1, Y_2\}$			
X2: N's + N *Bishop's Stortford*	P4: N's + N → N	S1: transliteration	Y2: N feminine singular *Bišopsstortforda*
$P_5=\{X_3, S_1, Y_2\}$			
X3: N + N's + N *St. Bishop's Town*	P5: N + N's + N → N	S1: transliteration	Y2: N feminine singular *Sentbišopsatauna*
$P_6=\{X_4, S_1, Y_2\}$			
X4: N + N *Bishop Auckland* *North Ronaldsay*	P6: N + N → N	S1: transliteration	Y2: N feminine singular *Bošopoklenda* *Nortronaldseja*
$P_7=\{X_5, S_1, Y_2\}$			
X5: A + N *South Ribble, Green Isle*	P7: A + N → N	S1: transliteration	Y2: N feminine singular *Sautribla* *Grīnaila*
$P_8=\{X_6, S_3, Y_4\}$			
X6: N + P + N *Milford upon Sea*	P8: N + P + N → N + P + N	S3: transliteration + translation	Y4: N feminine singular genitive + P + N

Stratford upon Avon			*Milforda pie jūras, Stradforda pie Avona*
$P_9=\{X_6, S_1, Y_5\}$			
X6: N + P + *Longville in the Dale*	P9: N + P + N → N + N	S1: transliteration	Y5: N feminine singular genitive + N feminine singular locative *Longvila Deilā*
$P_{10}=\{X_7, S_1, Y_2\}$			
X7: A + A + N *North East Coast*	P10: A + A + N → N	S1: transliteration	Y2: N feminine singular *Nortīstkosta*
$P_{11}=\{X_8, S_2, Y_3\}$			
X8: N + C + N *Sandal & Agbrigg*	P11: N + C + N → N + N	S2: transliteration + nomenclature word	Y3: N feminine singular genitive + N *Sendalendagbrigas stacija*
$P_{12}=\{X_4, S_3, Y_6\}$			
X4: N + N *Newton Point*	P12: N + N → N + N	S3: transliteration + translation	Y6: N masculine singular genitive + N *Ņūtona zemesrags*
$P_{13}=\{X_6, S_1, Y_1\}$			
X6: N + P + N *Horse of Copinsay*	P:13 N + P + N → N	S1: transliteration	Y1: N masculine singular *Horsofkopinsejs*
$P_{14}=\{X_7, S_3, Y_7\}$			
X7: N + N + N *Gog Magog Hills*	P14: N + N + N → N + N	S3: transliteration + translation	Y7: N masculine plural genitive +N *Gogmagogu kalni*

Table 1. Examples of English-Latvian Linguistic Toponym Translation Patterns

4 Evaluation and Limitations

The current MT evaluation theory and practice lacks in evaluation methods for toponym translation task. One of the reasons could be that it is not clear what the correct toponym translation is, since results may vary and more than one target toponymic unit is acceptable. As a result, scores calculated with a single target variant will underestimate translation accuracy. Moreover, human translations are often inaccurate as well.

Existing English-Latvian MT systems[1] do not implement any OOV algorithms to translate toponymic units. Thus, we had no possibility to compare our algorithm with other MT performance.

For evaluation purposes we compared translation results of our translation module with reference (human) translations from two bilingual dictionaries. 330 English toponymic units of different types with Latvian translation equivalents were manually extracted from dictionaries and processed with our OOV toponym translation module. We set the following evaluation scores:

- if the translation result coincides with the corresponding linguistic toponym translation pattern then the translation is *accurate* and the score is 1;
- if the translation result deviates from the corresponding linguistic toponym translation pattern then the translation is *inaccurate*, and the score is 0,5 for one distinction and 0 for more distinctions.

We accept variants as they were also described by linguistic toponym translation patterns (in transliteration rules). As a result, the accuracy of translation is 67% on the whole test set, 58% on the set containing one-word toponymic units, and 81% on multi-word test set.

[1] English-Latvian Pragma Expert: www.acl.lv, English-Latvian Google: http://translate.google.com, English-Latvian Tilde http://www.tilde.lv/English/portal/go/tilde/3777/en-US/DesktopDefault.aspx (November, 2008)

5 Conclusions and Future Work

We have described the pattern-based toponym translation approach developed for the English-Latvian language pair. The focus of the paper is on the detailed description of OOV toponym processing and describes possible translation strategies and linguistic toponym translation patterns with examples and evaluation results.

We can conclude that for the implemented rule-based approach there is much room for possible improvements, and evaluation results prove this statement. The main reason, why toponym processing is such a challenge for an MT task, is the necessity of knowledge of toponym rendering rules, variety of languages as well as a considerable amount of history and culture (Castañeda-Hernández, 2004). It is impossible to formalize this process completely and it is obvious that there can be mistakes in automated translation of toponymic units.

Corpus-based approach has not been applied in this research due to the lack of monolingual and bilingual linguistic resources. However, the issue of compiling a multilingual corpus of toponym-referenced texts for the Latvian language is being studied.

We consider the present research as the starting point for such tasks as multilingual cross-language MT of toponyms and application to other languages, especially Cyrillic or other non-Latin scripts.

Acknowledgement

We would like to thank Raivis Skadiņš for his comments and remarks on the article, Lars Ahrenberg for discussions on toponym machine translation, and Lars Borin for general discussions on toponymy.

This research was carried out in the framework of the project no. 045335 – TRIPOD project (TRI-Partite multimedia Object Description) co-funded by the European Commission within the Sixth Framework Programme.

References

Antonija Ahero. 2006. *English Proper Name Rendering into the Latvian Language* (Angļu Īpašvārdu Atveide Latviešu Valodā). Zinātne, Rīga.

Iñaki Alegria, Nerea Ezeiza, Izaskun Fernandez. 2006. Named entities translation based on comparable corpora. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Multi-word expressions in a Multilingual Context*, Italy. Pp.1-8.

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, USA. Pp.400-408.

Bogdan Babych and Anthony Hartley. 2003. Improving Machine Translation Quality with Automatic Named Entity Recognition. *Proceedings of the 7th European Association for Machine Translation Workshop Improving machine translation through other language Technology Tools*, Hungary. Pp.1-8.

Bogdan Babych and Anthony Hartley. 2004. Selecting Translation Strategies in MT using Automatic Named Entity Recognition. *Proceedings of the 9th European Association for Machine Translation Workshop Broadening horizons of machine translation and its applications*, Malta. Pp.18-25.

Gilberto Castañeda-Hernández. 2004. Navigating through Treacherous Waters: The Translation of Geographical Names. *Translation Journal*, 8(2): [electronic resource]: http://accurapid.com/journal/28names.htm#1

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2007. Collapsed consonant and vowel models: new approaches for English-Persian transliteration and back-transliteration. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Czech Republic. Pp.648-655.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics*, 24(4):599-612.

Chun-Jen Lee and Jason S. Chang. 2003. Acquisition of English-Chinese Transliteration Word Pairs from Parallel-Aligned Texts using a Statistical Machine Translation Model. *Proceedings of Human Language Technologies – The North American Chapter of the Association for Computational Linguistics Workshop: Building and Using parallel Texts Data Driven Machine Translation and Beyond*, Canada. Pp.96-103.

Geoffrey Leech. 1981. *Semantics. The Study of Meaning.* 2nd edition. Penguin, London, England, UK.

Jochen L. Leidner. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* PhD thesis. Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. *Proceedings of the 42nd Annual Meeting on association for Computational Linguistics*. Spain. Pp.159–166.

Katja Markert and Malvina Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, France. Pp.1385-1392.

Helen M. Meng, Wai-Kit Lo, Berlin Chen, and Karen Tang. 2001. Generate Phonetic Cognates to Handle Named Entities in English-Chinese cross-language spoken document retrieval. *Proceedings of Institute of Electrical and Electronics Engineers Automatic Speech Recognition and Understanding Workshop*, Italy.

Jong-Hoon Oh and Key-Sun Choi. 2002. An English-Korean Transliteration Model Using Pronunciation and Contextual Rules. *Proceedings of the 19th International Conference on Computational Linguistics*, Taiwan, 1:1-7.

Richard Sproat, Tao Tao, and Cheng-Xiang Zhai. 2006. Named entity transliteration with comparable corpora. *Proceedings of the 44th Annual meeting of the Association for Computational Linguistics*, Australia. Pp.73-80.

Bonnie Glover Stalls and Kevin Knight. 1998. Translating Names and Technical Terms in Arabic Text. *Proceedings of the Coling / Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages*, Canada. Pp.365-266.

Wolodja Wentland, Johannes Knopp, Carina Silberer, and Matthias Hartung. 2008. Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. *Proceedings of the 6th Language Resources and Evaluation Conference*, Morocco.

Min Zhang, Haizhou Li, and Jian Su. 2004. Direct Orthographical Mapping for Machine Transliteration. *Proceedings of the 20th International Conference on Computational Linguistics*, Switzerland.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206